

機械学習による 遺伝子型判定の性能向上

より正確な遺伝子型判定で信頼性の高い疾病リスク予測を目指す

近年、DNA（デオキシリボ核酸）マイクロアレイを用いた疾病リスクを予測するための遺伝子解析サービスが急激に広がっていますが、そのベースとなる遺伝子型判定については、改善の余地が残っています。信頼性の高い疾病リスクを算出するには、体質に起因する疾病のかかりやすさとの深いSNP（Single Nucleotide Polymorphism：一塩基多型）の遺伝子型を正確に判定することが重要です。

そこで東芝は、機械学習により、SNPにおける遺伝子型判定の性能を向上させる技術を開発し、性能改善の見込みを得ることができました。個別化医療の発展に向けて、この技術の貢献が期待できます。

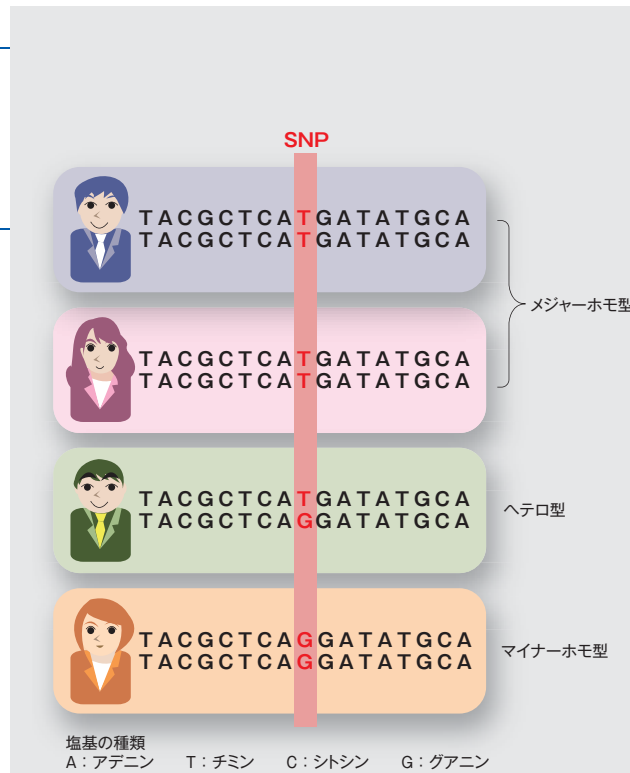


図1. SNPの遺伝子型の一例 — 2種類の遺伝子の組合せにより、メジャーホモ型、ヘテロ型、及びマイナーホモ型のいずれかの遺伝子型に割り当てられます。

SNPの遺伝子型と疾病リスク

人の遺伝子は、4種類の塩基の配列で表現されています。人は父方と母方の二つの遺伝子を受け継ぎますが、塩基配列のある特定の位置（遺伝子座）で父方の塩基と母方の塩基が異なる場合や、父母の塩基両方とも別の人とは異なる塩基を持つ場合があります。このような遺伝子座をSNPと呼びます。SNPには、基本的に、父方と母方のそれぞれから受け継いだ塩基の組合せにより、3種類の遺伝子型（メジャーホモ型、ヘテロ型、及びマイナーホモ型）が存在します（図1）。SNPにおける遺伝子型は先天的な体質に関連するとされており、その遺伝子型により、体質に個人差が生じ、疾病を発症するリスクが

異なることがあります。この仕組みを疾病リスク予測に活用するためには、なるべく多くのSNPの遺伝子型をできるだけ正確に判定する必要があります。

DNAマイクロアレイ

遺伝子座を特徴づける四つの塩基の中で、アデニン（A）とチミン（T）、シトシン（C）とグアニン（G）は互いに相補的な関係にあります。DNAマイクロアレイは、SNPの手前までの既知の遺伝子配列の相補的な塩基配列情報に基づいて設計したDNAプローブを面状に配列したチップです。このアレイを用いた遺伝子型判定では、血液などの検体からDNAを抽出した後、アレイ上のDNAプローブでSNPの手前までのDNAを捕え、その先端のSNPに対して相補的な

塩基を付加した蛍光色素を結合させます。蛍光色素の種類は塩基ごとに決まてあり、それぞれの蛍光色の強度を画像スキャナ装置で検出し（図2）、その信号を複数検体分まとめて、クラスタリングアルゴリズムを用いてSNPの遺伝子型を判定します。

遺伝子型判定技術

ここでの遺伝子型判定とは各SNPに対して2種類の蛍光強度の信号データの比と和の2軸で表された特徴空間上で各検体の値をプロットし、そこで形成される3クラスをクラスタリングアルゴリズムで決定し、前述の三つの遺伝子型に対応付けることを指します。判定技術の従来方式では、事前に各遺伝子型がどのあたりに分布するかの確率

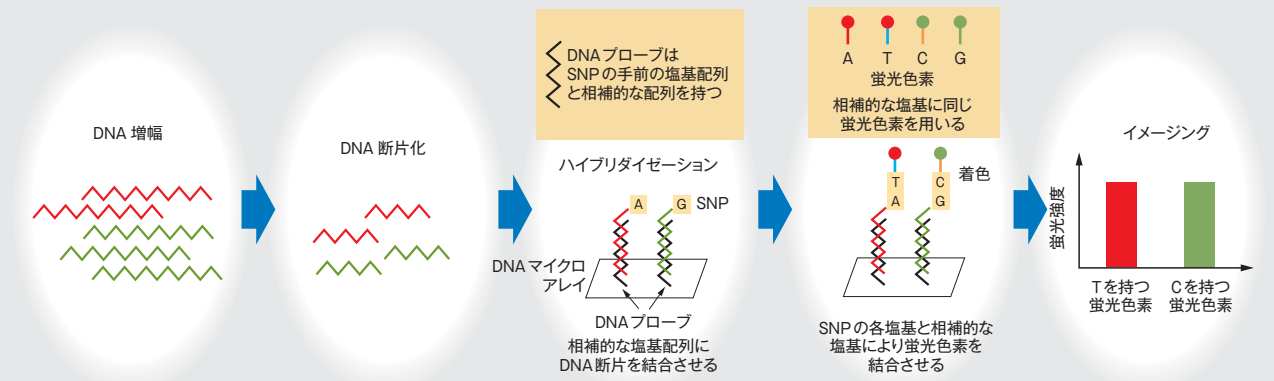


図2. 検体のDNA増幅から蛍光強度の信号データを得るまでの流れの例 — 血液などからDNAを抽出後、増幅、断片化、ハイブリダイゼーション、着色、及びイメージングを経て、蛍光強度の信号データを取得します。

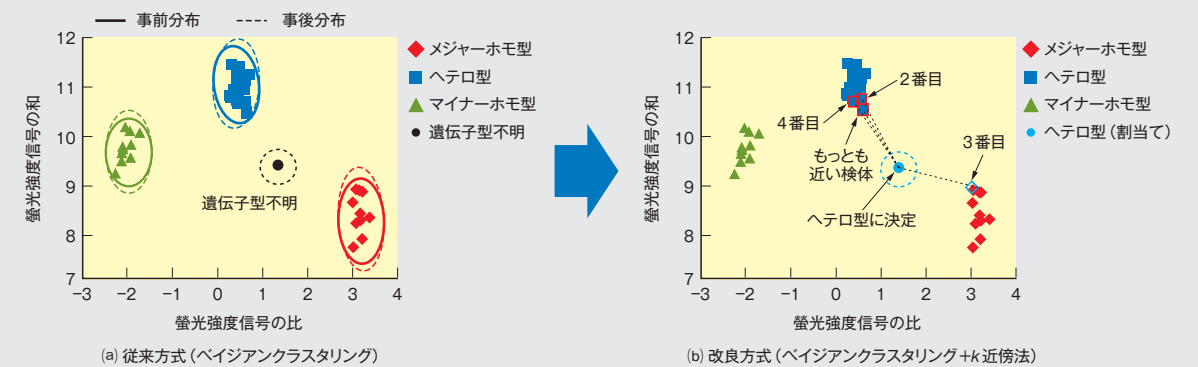


図3. 遺伝子型判定 — 従来方式(a)では遺伝子型判定結果は事前及び事後分布に大きく依存しますが、改良方式(b)では、k近傍法により、事前分布から外れた検体の型判定を行います。

と、実際の検体がどこに分布するかを表す座標に基づいてもっともらしい遺伝子型を判定するベイジアンクラスタリングに基づく方法⁽¹⁾を用います。ところが、いずれのクラスタからも離れている少数の検体（図3(a)）では遺伝子型を必ずしも正しく判定することができません。そこで東芝は、この判定が困難な検体に対して、既に判定結果を得られている検体との距離を計算し、この距離が小さい順に上位k検体を選定して、これらの中で多数を占めるクラスターの遺伝子型を割り当てるk近傍法をベイジアンクラスタリングに加えた方式を提案しました（図3(b)）。

判定性能の評価

国際研究協力の一つである「1000

人ゲノムプロジェクト」の公開SNPデータのうち、ヨルバ族のベンチマークデータ（数十検体を対象とした約17万種類のSNP）を用いて、従来方式と改良方式を検出率で評価したところ、従来方式よりも1%程度の改善を確認できました。これは約1,700種類のSNPに相当し、一つでも多くのSNPを正しく判定することが重要な疾病関連の分野において、短期間で改良の第一歩を踏み出したと言えます。

今後の展望

遺伝子型判定の性能向上に向けた技術開発は緒に就いたばかりですが、今後も、個別化医療の推進の一端を支えるために、様々なアイデアを活用し、更なる改良を進めていきます⁽²⁾。

文献

- Affymetrix®, "DMET™ Plus genotyping and copy number methods". Affymetrix homepage. <http://media.affymetrix.com/support/technical/whitepapers/dmet_plus_algorithm_whitepaper_v1.pdf>. (accessed 2015-07-15).
- Fukushima, A. et al. "A Proposal for Improvement of Genotyping Performance for Ethnically Homogeneous Population Using DNA Microarray". Proceedings of the 37th International Conference of the IEEE Engineering in Medicine and Biology Society, Milano, Italy, 2015-08, IEEE. (to be published).

植野 研

研究開発統括部
研究開発センター
システム技術ラボラトリー主任研究員